

R at UBER

From an SSA Central Operations perspective

Marc van Heerden

Analytics Manager, Sub-Saharan Africa

SATRDAY Cape Town

February 2017



UBER



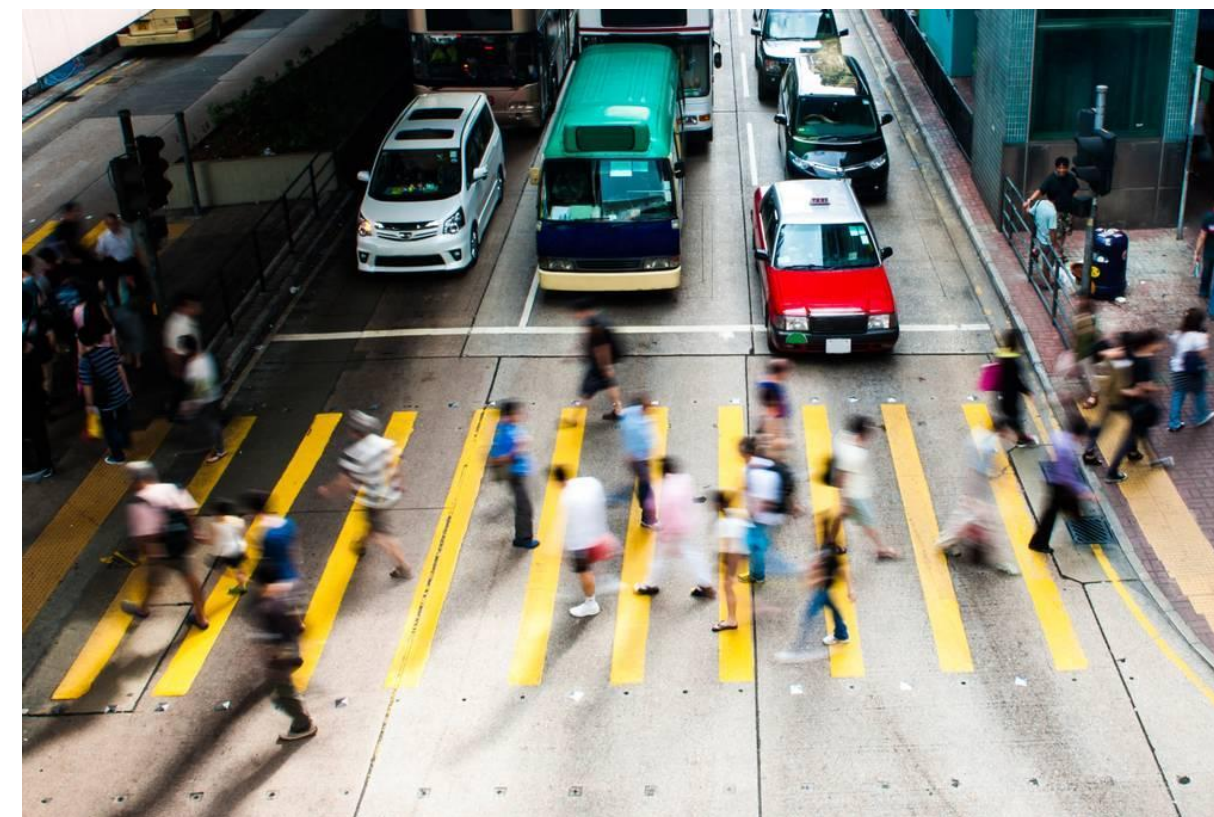
Business needs

Why R is a great tool for UBER



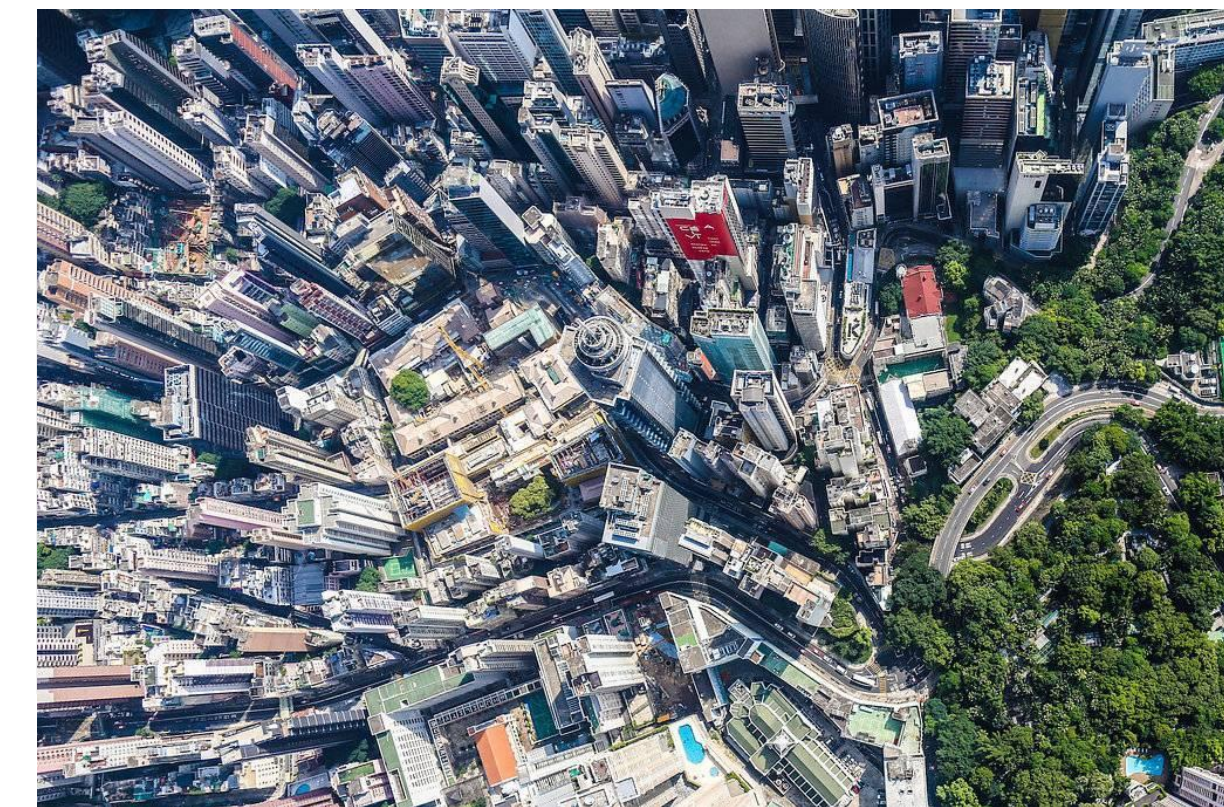
Reproducibility

Uber operates in various markets at different life stages. Reproducibility helps teams share knowledge, reduce duplication of effort and achieve greater consistency. Rmarkdown allows for reproducibility within the organisation.



Speed

Lean teams with competing priorities need to be able to quickly analyse data and set up reports. The Tidyverse ecosystem of packages allows for rapid, tidy data analysis.



Integration

We need a tool that can fetch data from various systems and send standardised output to different end points with little overhead. Internal packages allows us to meet this need.

Frequently used packages

ubeR

Making life at Uber easier one function at a time.

Andrew Stanek

Query various data sources

- No credential or odbc driver hassles
- Asynchronous
- Cache queries
- With parameters

Other features

- Easily move data to and from storage locations
- Set up piper contracts
- Create default project structure

shinydashboard.uber / uberplot

adding UBER styling to existing packages

shinydashboard.uber

- Winston Chang's shinydashboard package with uber themes included
- Ensures uniform style of shiny dashboards

uberplot

- ggplot2 themes and colour palettes
- Allows easy alignment

The tidyverse

ggplot2, tibble, tidyr, readr, purrr, and dplyr

Hadley Wickham et al

googlesheets

Manage Google Spreadsheets from R Jennifer Bryan and Joanna Zhao

why use google sheets

- Allows for less strict access control for external partners or contractors
- Integration with google forms
- Great for collaboration
- Some people really like spreadsheets
- Temporary solution

how googlesheets helps

- allows for automated updating which results in saved time, better accuracy and better recency.
- allows for larger data sets to be moved to google sheets without failure

xgboost

Extreme Gradient Boosting

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich and Yuan Tang

fast, accurate models

- runs in parallel with no setup required
- great feature importance plots

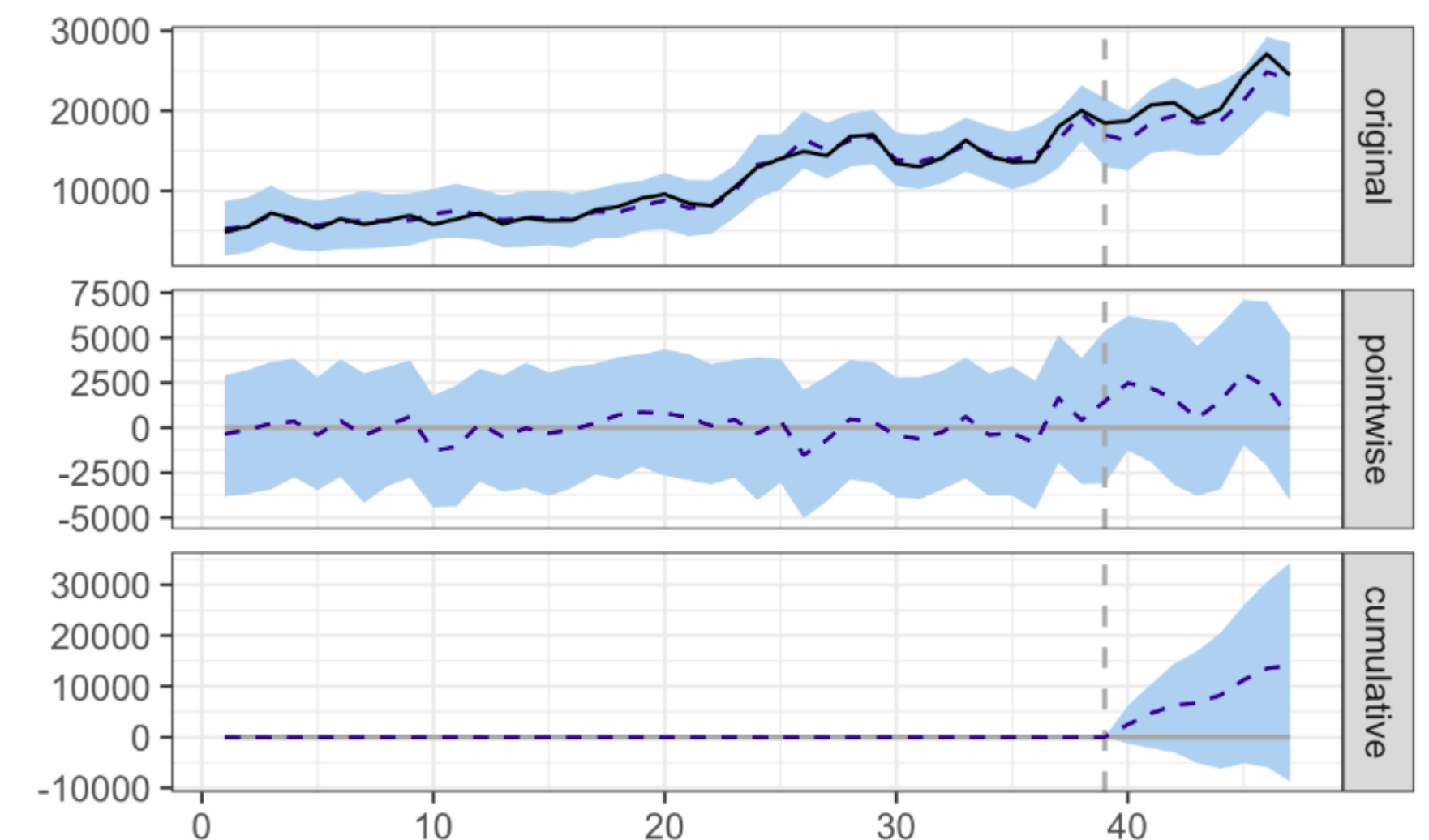
CausalImpact

Inferring causal impact using Bayesian structural time-series models

Brodersen et al

Quick inference where control groups are not an option

- Inability to treat partners or riders differently means that often control groups are not possible
- City-wide marketing also has no control group
- Quickly provides insight into the effect of an intervention on a city's metrics by comparing to a group of correlated cities.
- Provides easy-to-communicate visualisations



Supporting infrastructure

Shiny and ad hoc servers

Multiple servers to host Shiny Apps, run cron jobs and access production APIs

An easy, safe way to share information and run tasks

- Shiny servers host dashboards within existing access control
- HTML folder for sharing reports within existing access control
- Ad hoc servers provide access to APIs used for administration
- Set up cron jobs to automatically populate google sheet reports

Piper

Piper is a data workflow management platform for all of Uber. Piper enables users to author, manage and execute data workflows on a single platform.

A more reliable way to run regular queries

- Runs scheduled queries and stores the results
- Automatic retry on failure
- Allows for resource allocation
- Provides data for Shiny apps and other regular reports

Data Science Workbench (Beta)

Unleash the productivity of the Data Science community at Uber by providing scalable infrastructure, tools, customization and support.

A web based hub for all data scientists at Uber

- Easy access to fully configured and managed RStudio Server and Jupyter Notebooks
- A way to save, view, and share scripts and files
- Each session has a highly-available file system mounted within it. Any work performed in a session is reliably saved. You can view these files and share them in the Files view of DSW
- Integrated with Uber's data ecosystem
- Aim to improve productivity by 10-15%
- A central hub where data scientists can find and reuse knowledge



Sessions List

+ CREATE NEW

Create new session

Complete this form so that we can allocate resources and prepare your new session.

Session Name

MARCV_SESSION

Session Type

RSTUDIO

Instance Size

MEDIUM

CLOSE

CREATE

An unusual use case

Process automation in R

Interim solutions for automating business processes

Extensive internal support for R makes the case for using R instead of Python for automating bulk tasks.

- Managing corporate partnership administration
- Reducing load on in-person support by automating common tasks

Do you have any examples of how your organisation supports the use of R?

Do you have any unorthodox use cases for R?

Any questions?

Thank you

Proprietary and confidential © 2016 Uber Technologies, Inc. All rights reserved. No part of this document may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage or retrieval systems, without permission in writing from Uber. This document is intended only for the use of the individual or entity to whom it is addressed and contains information that is privileged, confidential or otherwise exempt from disclosure under applicable law. All recipients of this document are notified that the information contained herein includes proprietary and confidential information of Uber, and recipient may not make use of, disseminate, or in any way disclose this document or any of the enclosed information to any person other than employees of addressee to the extent necessary for consultations with authorized personnel of Uber.

A woman with curly hair, wearing a white shirt, blue overalls, and a brown bag, is walking across a city street. The background shows tall buildings, trees, and a clear blue sky. The Uber logo is overlaid on the bottom left of the image.

UBER